Scanning Human Gene Deserts for Long-Range Enhancers

Marcelo A. Nobrega,[1,2]* Ivan Ovcharenko,[1,2]*† Veena Afzal,[1,2]
Edward M. Rubin[1,2]‡

Approximately 25% of the genome consists of gene-poor regions greater than 500 kb, termed gene deserts (*1*). These segments have been minimally explored, and their functional significance remains elusive. One category of functional sequences postulated to lie in gene deserts is gene regulatory elements that have the ability to modulate gene expression over very long distances (*2*).

Human *DACH*, a gene expressed in numerous tissues and involved in the development of brain, limbs, and sensory organs (*3*, *4*), spans 430 kb and is bracketed by two gene deserts 870 kb and 1330 kb in length. A paucity of regulatory sequences has been identified in the proximity of the *DACH* promoter (*5*), suggesting that distal sequences, which could reside anywhere in a sea of sequence greater than 2630 kb, are likely responsible for the gene's complex expression characteristics.

To identify evolutionarily conserved footprints corresponding to putative *DACH* enhancers, we compared the human *DACH* sequence and the bracketing gene deserts to orthologous intervals in vertebrate species (Fig. 1A). Human and mouse sequence comparisons revealed a similar genomic structure within this region and identified 1098 conserved noncoding sequences (>100 bp and with >70% identity) in the 2630-kb targeted interval. To identify those with a greater likelihood of containing biological activity (*6*), we determined which of the human-mouse conserved sequences were also present in distant vertebrates, including frog, zebrafish, and two pufferfish (*1*). This decreased the number of conserved sequences to 32 (Fig. 1B).

To examine the possibility that these sequences, conserved over 1 billion years of parallel evolution, might represent enhancers, we explored their in vivo ability to drive gene expression with the use of a reporter assay system in transgenic mice. Nine elements were tested, representing a sampling of elements present in the two gene deserts and *DACH* introns, spread over a 1530-kb region surrounding the human *DACH*'s TATA box. Each corresponding human element was individually cloned upstream of a mouse heat shock protein 68 minimal promoter coupled to β-galactosidase and injected in fertilized mouse oocytes (*7*). Seven elements were shown to reproducibly drive β-galactosidase expression in a distinctive set of tissues in transgenic mice, recapitulating several aspects of *DACH* endogenous expression (Fig. 1C) (*3*, *4*).

Whereas the synteny of the orthologous noncoding elements flanking *DACH* is maintained in mammals and fish, the genes flanking *DACH* in these vertebrates differ (Fig. 1A). The failure of this chromosomal rearrangement to disturb the linear relation between the conserved noncoding elements and *DACH* further supports a functional relation between these sequences.

The demonstration that several of the enhancers characterized in this study reside in gene deserts highlights that these regions can indeed serve as reservoirs for sequence elements containing important functions. Moreover, our observations have implications for studies aiming to decipher the regulatory architecture of the human genome, as well as those exploring the functional impact of sequence variation. The size of genomic regions believed to be functionally linked to a particular gene may need to be expanded to take into account the possibility of essential regulatory sequences acting over near-megabase distances.

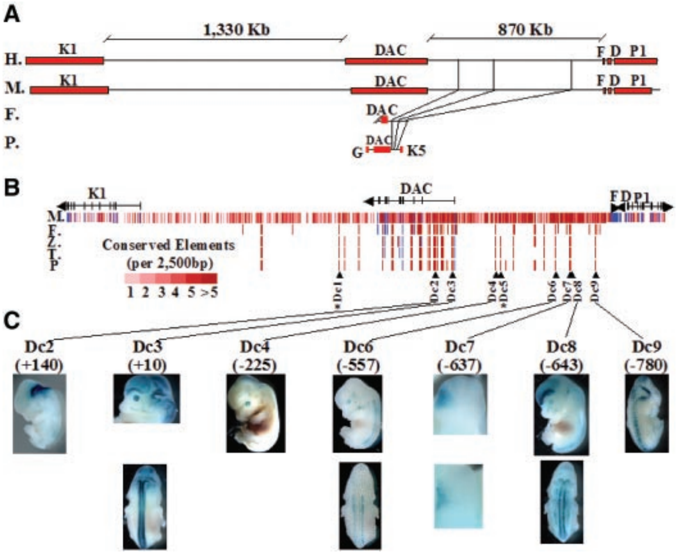

Fig. 1. (**A**) *DACH* locus in humans, mice, frog and pufferfish. Lines linking each panel represent positions of orthologous sequences. Genes are represented by their RefSeq name: DAC, DACH; K1, KLHL1; F, FLJ22624; D, DIS3; P1, PIBF1; G, GPR-18; K, KLF5. H, human; M, mouse; F, Frog; P, *Fugu rubripes*. (**B**) Sequence conservation plots (alignments were obtained at www-gsd.lbl.gov/vista). Bars correspond to sequence similarities between human and the species displayed. Blue bars denote exons; red bars denote noncoding sequences. Gradients of red indicate the number of conserved elements within 2500 bp windows. Asterisks denote elements with no detectable enhancer activity in this developmental stage. Z, zebrafish; T, *Tetraodon nigroviridis*. (**C**) Transgenic expression results. The distance (in kb) between each element and the human *DACH* TATA box is given in parenthesis. Expression patterns from representative 12.5 and 13.5 days post coitum mouse embryos are illustrated. Three or more independent transgenic founders were generated for each element.

References and Notes

1. Materials and methods are available as supporting material on *Science* Online.
2. L. A. Lettice *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **99**, 7548 (2002).
3. X. Caubit *et al.*, *Dev. Dyn.* **214**, 66 (1999).
4. R. J. Davis *et al.*, *Dev. Genes Evol.* **209**, 526 (1999).
5. O. Machon *et al.*, *Neuroscience* **112**, 951 (2002).
6. N. Ghanem *et al.*, *Genome. Res.* **13**, 533 (2003).
7. R. Kothary *et al.*, *Nature* **335**, 435 (1988).
8. We thank I. Plajzer-Frick and J. M. Collier for technical assistance and B. Black for the *hsp68/LacZ vector*. Supported by the National Heart Lung and Blood Institute Programs for Genomic Application (grant HL66728) and the U.S. DOE (contract no. DEAC0376SF00098).

Supporting Online Material
www.sciencemag.org/cgi/content/full/302/5644/413/DC1
Materials and Methods
Table S1

23 June 2003; accepted 8 September 2003

[1]U.S. Department of Energy Joint Genome Institute, Walnut Creek, CA 94598, USA. [2]Genome Sciences Department, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA.

*These authors contributed equally to this work.
†Present address: Biology and Biotechnology Research Program, Lawrence Livermore National Laboratory, Livermore, CA 94550, USA.
‡To whom correspondence should be addressed. E-mail: emrubin@lbl.gov